Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Radiologists remember mountains better than radiographs, or do they?

Karla K. Evans
Edith M. Marom
Myrna C. B. Godoy
Diana Palacio
Tara Sagebiel
Sonia Betancourt Cuellar
Mark McEntee
Charles Tian
Patrick C. Brennan
Tamara Miner Haygood

SPIE

Radiologists remember mountains better than radiographs, or do they?

Karla K. Evans,[a,*] Edith M. Marom,[b] Myrna C. B. Godoy,[b] Diana Palacio,[c] Tara Sagebiel,[b] Sonia Betancourt Cuellar,[b] Mark McEntee,[d] Charles Tian,[b] Patrick C. Brennan,[d] and Tamara Miner Haygood[b]

[a]The University of York, Department of Psychology, Heslington, York YO105DD, United Kingdom
[b]University of Texas MD Anderson Cancer Center, Department of Diagnostic Radiology, Unit 1475, 1515 Holcombe Boulevard, Houston, Texas 77030, United States
[c]The University of Arizona, College of Medicine, Department of Medical Imaging, 1501 North Cambell Avenue, Tucson, Arizona 85724-5067, United States
[d]University of Sydney, Medical Image Optimisation and Perception Group, Discipline of Medical Radiation Sciences (C42), Room M221, Cumberland Campus, Sydney NSW 2141, Australia

Abstract. Expertise with encoding material has been shown to aid long-term memory for that material. It is not clear how relevant this expertise is for image memorability (e.g., radiologists' memory for radiographs), and how robust over time. In two studies, we tested scene memory using a standard long-term memory paradigm. One compared the performance of radiologists to naïve observers on two image sets, chest radiographs and everyday scenes, and the other radiologists' memory with immediate as opposed to delayed recognition tests using musculoskeletal radiographs and forest scenes. Radiologists' memory was better than novices for images of expertise but no different for everyday scenes. With the heterogeneity of image sets equated, radiologists' expertise with radiographs afforded them better memory for the musculoskeletal radiographs than forest scenes. Enhanced memory for images of expertise disappeared over time, resulting in chance level performance for both image sets after weeks of delay. Expertise with the material is important for visual memorability but not to the same extent as idiosyncratic detail and variability of the image set. Similar memory decline with time for images of expertise as for everyday scenes further suggests that extended familiarity with an image is not a robust factor for visual memorability. [DOI: 10.1117/1.JMI.3.1.011005]

Keywords: visual episodic memory; perceptual expertise; memorability; radiographs; observer performance studies.

Paper 15139SSR received Jul. 9, 2015; accepted for publication Sep. 25, 2015; published online Nov. 3, 2015.

1 Introduction

Humans have a massive and high-fidelity visual long-term memory,[1] far superior to verbal memory.[2] Visual memory also highly supersedes auditory memory even for musicians.[3,4] We are well prepared to commit to memory the visual images of the sorts of scenes that we encounter in the world, and this ability raises a series of questions that are relevant for clinical research and practice in radiology such as robustness of that memory over time and the characteristics of the formed memory trace.

Investigators who design visual performance studies with radiologists as the intended participants take care to negate as much as possible the effects of memory for the images being shown on the outcome of the study. Some of these measures require little time or effort, for instance, showing images in different random orders each time. Commonly, investigators also build in a time gap between viewings. This can become lengthy, and there is little guidance in the literature as to how long it should be or whether such a time gap is needed at all.

The reason it is not clear if a time gap is needed is that little is known about the degree to which radiologists recognize specific radiologic images. Mnemonic ability for representative stimuli from a domain of expertise has been seen as essential for acquisition of that expertise, and studies have shown that memory for images can be augmented by expertise in the field to which the images relate. Master chess, bridge, or sports players and computer programmers have superior ability compared to nonexperts in memorizing meaningful material from their general domains of expertise,[5–10] but not for randomly rearranged versions of those stimuli. This memory seems to be linked to recognition of specific patterns. For example, there are only a certain number of ways that chess pieces are likely to be configured on a board, and a chess master can recognize one of these patterns when it is seen again. Radiologists, however, may rely less on the ability to remember specific meaningful arrangements than on learning the patterns that may signify a specific disease with the hope that they can then recognize these patterns even when they vary in their appearance from one patient to the next, and this might then help them with diagnosis. The degree to which this ability may be associated with an ability to recognize specific radiographs is unclear.

Two studies that have focused on visual recognition memory in breast imaging found that radiologists did not recall images that they had earlier interpreted when mixed with mammograms that had been interpreted by others,[11] and their absolute performance in recognizing previously encountered mammograms was quite

*Address all correspondence to: Karla K. Evans, E-mail: karla.evans@york.ac.uk

2329-4302/2015/$25.00 © 2015 SPIE

poor even though superior to that of nonexperts, and far worse than their performance with everyday scenes and objects.[12] In studies with chest radiographs, there are findings of superior memory for experienced radiologists compared to first-year residents for chest radiographs with abnormalities[13–15] and weak incidental memory for repetition of chest images with abnormalities after short intervals.[16] Thus, there is no clear evidence that radiologists have a massive memory for representative radiographic images.

The main motivation for the present investigation was to address the aforementioned concern for observer studies in radiology in regard to the effects of visual memory for the study material. Therefore, the purpose of the present studies was fourfold. First, we wished to build on the studies mentioned and further investigate the relationship between visual recognition memory and perceptual expertise by comparing radiologists' and naïve subjects' recognition memory for chest radiographs versus everyday scenes. We chose chest radiographs because (1) they come close to being the medical equivalent of scenes, in that they are composed of several different structures (e.g., bones, lung, heart, vessels, and the silhouette of the outer surface of the body) with different very specific spatial layouts, and (2) they demonstrate less homogeneity than mammograms. To build on the results of this first experiment, we then wanted to test radiologists' recognition memory with datasets that would "even the playing field" as compared with the everyday scene versus chest radiograph comparison. For this part of our study, we chose everyday scenes from just one class of image (forests) and compared them with a more varied assortment of musculoskeletal radiographs. Third, we wanted to investigate the robustness of visual recognition memory for radiologists over time. We did this by combining in the second experiment tests of both immediate and delayed memory. Finally, we wanted to investigate the degree to which expert radiologists can predict which radiographs will be easy or difficult to recognize. By choosing these questions, we believed we could move knowledge forward in several directions with just two interconnected experiments.

2 Experiment 1

2.1 *Methods*

2.1.1 *Study participants*

To evaluate visual recognition memory of medical experts for images in their general domain of expertise, we used a standard procedure from psychological sciences. Two groups of participants took part in the present study, a group of radiologists and a control group of medically naïve participants. The expert group consisted of 12 board-certified radiologists (six males and six females, 6 to 39 years of experience after residency), not all subspecialists in chest radiology yet reading on average 140 chest radiographs per week. The control group of 12 medically naïve observers (seven females and five males) had no medical background and an age range from 21 to 55 years. Informed consent was obtained from participating radiologists and medically naïve participants.

2.1.2 *Stimuli and apparatus*

A radiologist (XX), who did not later participate as an observer, obtained 108 anonymized chest radiographs from the University of Texas, MD Anderson Cancer Center. Informed consent was waived with respect to patients whose radiographs were used. To avoid the bias that might come of having a hand-selected group of radiographs, these 108 radiographs were the posterior–anterior (PA) projections associated with 108 consecutive outpatient PA and lateral chest radiographs that she encountered in her clinical practice. They were a mixture of images with and without abnormalities. No medical history data were associated with them as shown to the observers. The radiologist who collected them also indicated whether an abnormality was present or absent [Fig. 1(b)]. The stimulus set used to test memory for real scenes consisted of 108 real photographs of different categories of images (e.g., beach, mountain, cityscape, forest, and room interior) obtained from a public image dataset hosted by the Computational Visual Cognition Laboratory at MIT[17] [Fig. 1(a)].

The experiment with the medically naïve group was conducted on a Macintosh computer running MacOS X. The experiment for the radiologists was run on a Dell Precision M6500 computer (Austin, Texas). Both computers were controlled by MATLAB® 7.5.0 and the Psychophysics Toolbox, Version 3.[18,19]

2.1.3 *Ranking of chest radiographs*

To allow us to test how well it can be predicted which radiographs will be recognized and which will not, we also placed all 108 images in order in a PowerPoint program and requested three board-certified radiologists, all with subspecialty expertise in thoracic imaging, to divide the images into three equal groups: those they thought would be easy to recognize, difficult to recognize, and of intermediate difficulty. None of these three radiologists participated as an observer. All were encouraged to apply whichever criteria seemed appropriate to each individual image to determine ease or difficulty of recognition.

2.1.4 *Procedure*

This prospective study was reviewed and approved by the University of Texas, MD Anderson Cancer Center, institutional review board and was HIPAA compliant. The study was composed of study and test phases for each of two stimulus sets. The study and test phases were done back to back for one stimulus set before the subject went on to the next set. In the study phase, each participant saw 72 images that were randomly taken from the 108 chest radiographs or everyday scenes. The study images were consecutively presented on the computer display, each for 3 s with no time between the images, resulting in a total time of 3 min and 36 s for the study phase. Participants were told to memorize the images in preparation for a recognition test. The test phase followed immediately after the study phase. In the test phase, participants saw a sequence of 72 images, of which 36 were randomly chosen old images from the study phase, and the remaining 36 were completely new images.

Each test image was presented one at a time on the display until the participant responded. Participants were asked to label each image as "old" or "new" by pressing the appropriate computer key. The images remained on the screen until the response was given, and immediate feedback was provided for each test image. All participants completed the test and study phase blocks for the two image types for a total of 144 test trials. The order of the blocks was counterbalanced across participants. Our principal unit of analysis was the probability of a hit minus the probability of a false alarm (hits – false alarms), which we refer to as recognition accuracy. We also report performance in terms of percentage correct and assess differences using the

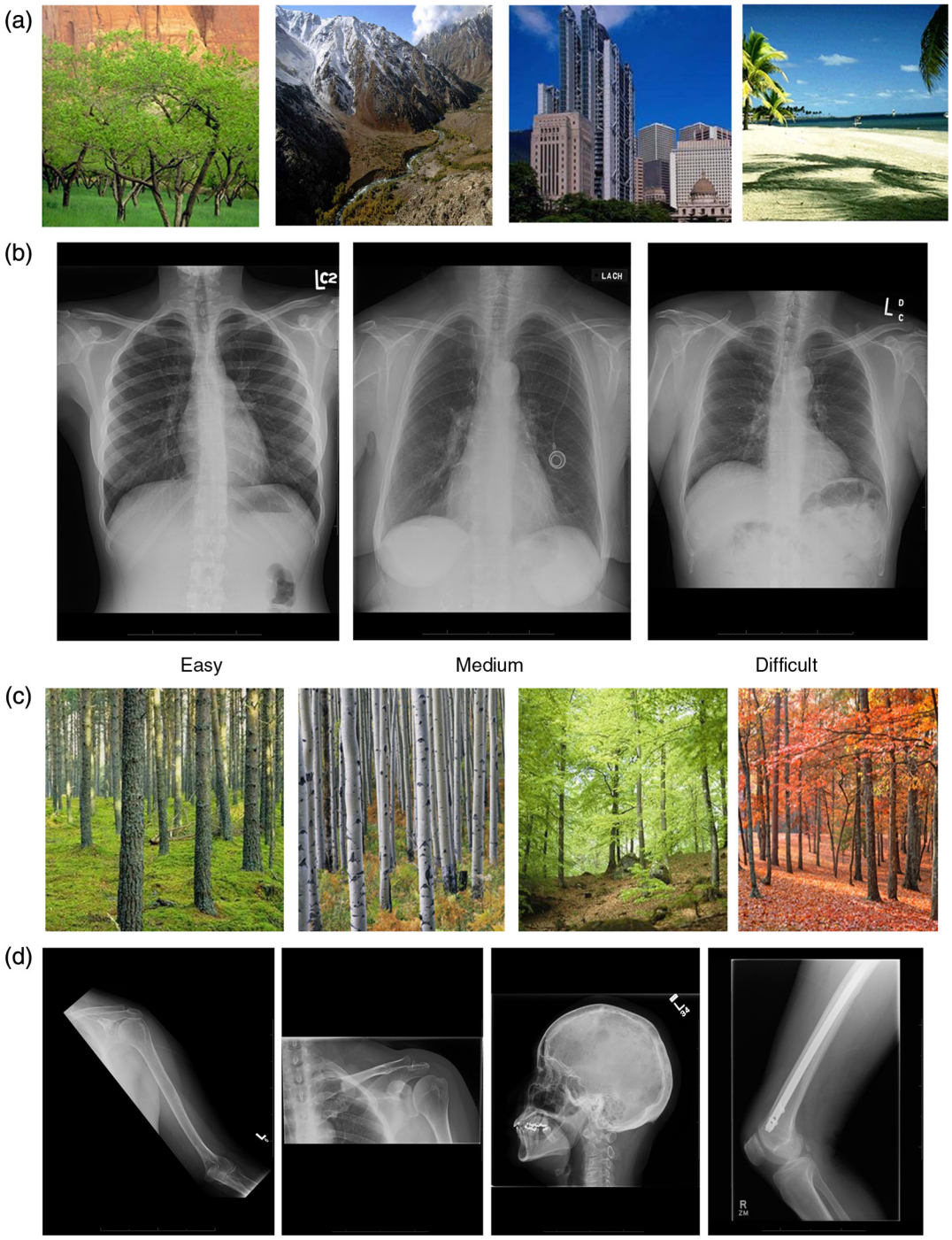

Fig. 1 Examples of images used to test visual recognition memory of radiologists and the naïve control group. (a) Four examples of real scenes used as stimuli, each one representative of one of the scene categories (forest, mountain, cityscape, and beach). (b) Three examples of chest radiographs used as stimuli, each one representative of one of the three levels of memorability (easy, medium, and difficult to remember). (c) Examples of a homogeneous set of natural scenes (forest) used in Experiment 2. (d) Example musculoskeletal radiographs used in Experiment 2.

signal detection measures of memory sensitivity (d′) and response criterion (c).

3 Results

3.1 *Recognition Memory*

The aim of this prospective study was to examine visual recognition memory of radiologists in general and more specifically for images from their general domain of expertise (chest radiographs). In addition, we wanted to see how they compare to medically naïve participants. Based on the results of a mixed model analysis of variance (ANOVA), both groups were very good at remembering everyday scenes (radiologists 85%, s.e.m. = 1.6%, $d' = 2.06$; naïve 81%, s.e.m. = 1.9%, $d' = 1.77$) and significantly worse for remembering chest radiographs [radiologists 65%, s.e.m. = 2.1%, $d' = .80$; naïve 55%, s.e.m. = 1.8%,

$d' = .24$; $F(1,22) = 164.7$, $p < 0.0001$] but still significantly above chance for radiologists (radiologists $p < .0001$; naïve $p < .026$). Though the naïve participants' memory performance for radiographs was above chance, it was very poor, with only 55% correct.

When testing memory for real photographs of everyday scenes, there is no significant difference in memory performance between radiologists and the naïve group [$t(22) = 1.61$, $p = .122$; see Figs. 2(a) and 2(b)]. The situation is quite different for chest radiographs. Radiologists are significantly better at remembering chest radiographs than naïve observers [($t(22) = 3.83$, $p < .001$; see Figs. 2(a) and 2(c)]. In Figs. 2(b) and 2(c), we have replotted our findings as scatter plots of z score hits against z score false positives by image type and group since this permits us to normalize the scores to a central mean, thus allowing a comparison of measures with very different ranges of absolute values. Compared in this way, we also see that radiologists show better memory for the images from their general domain of expertise [Fig. 2(c)] in comparison to the control group but no difference when visual stimuli are real scenes [Fig. 2(b)].

3.1.1 *Correlation with memorability scoring*

First, we looked at the level of agreement between the three board-certified radiologists with subspecialty expertise in thoracic imaging. We examined their ranking scores for individual images and found that on average at least two radiologists agreed 90% of the time. All three radiologists agreed for 77 radiographs (71.3%), and we consider this to be a ranking with consensus. Their rankings also positively correlate ($r = 0.69$ for all images; $r = 0.79$ for images with consensus) with the presence of an abnormality, with radiographs with no abnormalities being rated as more likely to be hard to recognize than those with abnormalities. However, we found no significant correlation between the scoring of memorability and the actual readers' performance on the memory test, either for all images ($r = -0.15$; percentage of agreement on easy = 68%; medium = 67%; difficult = 61%) or for images with consensus (r = –0.22, easy = 69%; medium = 67%; difficult = 61%). Thus, the images that were ranked as easy to remember by independent raters were not remembered significantly better than other images by radiologists who participated in the experiment.

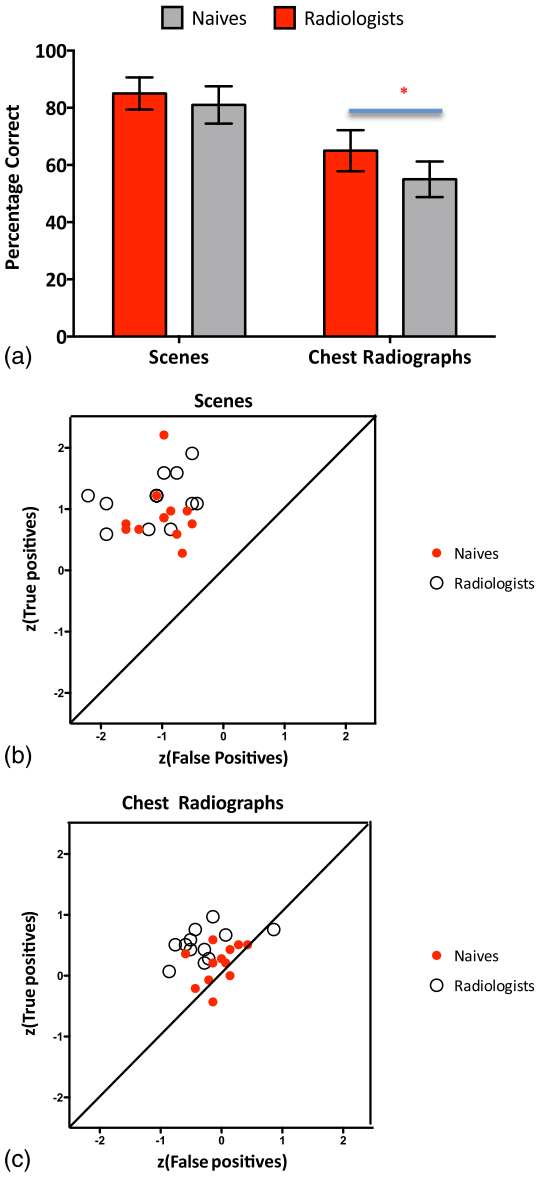

Fig. 2 Performance on visual recognition memory test of the radiologists and medically naïve participants for two real scenes and chest radiographs. (a) Average accuracy for the two groups across two different image sets. Error bars represent the standard error of the mean. An asterisk signifies a significant statistical difference. (b) Scatter plot of z score true positives against z score false positives by group for chest radiographs (radiologists' average: 69% hits, 38% false positives; medically naïve average: 58% hits, 48% false positives). (c) Scatter plot of z score true positives against z score false positives by group for scenes (radiologists' average: 86% hits, 17% false positives; medically naïve average: 80% hits, 17% false positives).

4 Experiment 2

The intention of the second experiment was to determine (1) the degree to which the modest memory we found for chest radiographs may be improved by using a wider variety of images, (2) the degree to which the fairly robust memory we found for everyday scenes in Experiment 1 may be degraded by using a more homogeneous type of image, and (3) the degree to which the memory of each is degraded by passage of a few weeks, the time period in question being chosen because it is a reasonable estimate of the time lapse actually used in many radiology projects.

4.1 *Methods*

4.1.1 *Study participants*

The second experiment only involved a group of radiologists. We compared experts' performance on a more heterogeneous set of radiographs (musculoskeletal) and a more homogeneous set of one category of natural scenes (forests) when their memory was probed immediately and with a delay of approximately 7 weeks. The expert group consisted of 11 American College of Radiology board-certified attending radiologists (four males and seven females, 4.5 to 38 years of experience after residency). One subspecialized in musculoskeletal imaging, six in thoracic imaging, and four in abdominal imaging. They practice in a large academic hospital, interpreting on average 300 imaging studies per week. All participants gave informed consent to participate in this prospective study.

4.1.2 *Stimuli and apparatus*

A radiologist (XX), who did not later participate as an observer, obtained 216 musculoskeletal radiographs from the University of Texas, MD Anderson Cancer Center. Informed consent was waived with respect to patients whose radiographs were used. To decrease the bias that might come of having a hand-selected group of radiographs, these 216 radiographs were taken from 216 consecutive patients with musculoskeletal radiographs encountered in clinical practice. As it was the intention to make this a varied set of radiographs, XX then chose just one image from among all the musculoskeletal radiographs that had been performed that day for each patient, varying the body part that was imaged and the projection to maximize variation of musculoskeletal radiographs (Table 1). All musculoskeletal radiographs were anonymized, and there were no medical history data associated with them when shown to observers [Fig. 1(d)].

The stimulus set used to test memory for real scenes consisted of 216 real photographs of only one image category, forests, obtained from a public image dataset hosted by the Computational Visual Cognition Laboratory at MIT[17] [Fig. 1(c)]. The set of forests was quite varied, with coniferous, palm, deciduous, dense or sparse forests, and orchards. Forests were photographed in every season. In the opinions of the investigators, all images were distinguishable from one another.

Table 1 Characteristics of radiographs used in Experiment 2

Body part imaged	Total	AP views	Lateral views	Oblique views	Specialty views	Type of specialty view, if any
Ankle	4	1	1	2	0	
Cervical spine	14	5	6	1	2	Flexion
Clavicle	4	4	0	0	0	
Elbow	2	0	2	0	0	
Femur	20	8[a]	12	0	0	
Foot	6	3	1	2	0	
Forearm	9	8	1	0	0	
Hand	4	1	2	1	0	
Hip	13	7	6[b]	0	0	
Humerus	11	11	0	0	0	
Knee	16	9	6	1	0	
Lumbar spine	19	11	7	0	1	Extension
Mandible	1	1	0	0	0	
Pelvis	16	16	0	0	0	
ribs	19	3	0	11[c]	5	Low AP
Sacrum	3	0	3	0	0	
Scapula	1	0	1	0	0	
Shoulder	20	15[d]	0	0	5	Y
Skull	9	0	9	0	0	
Tibia and fibula	12	9[e]	3	0	0	
Thoracic spine	9	5	4	0	0	
thoraco-lumbar spine	1	0	1	0	0	
Wrist	3	2	1	0	0	
	216					

[a]Five AP views of the proximal femur and three AP views of the distal femur.
[b]Frog-leg lateral views.
[c]Seven right posterior oblique and four left posterior oblique.
[d]One straight AP, seven AP in internal rotation, and seven AP in external rotation.
[e]Three APs of the whole tibia and fibula and three each of just the proximal and distal parts.

The experiment was run on a Dell Precision M6500 computer (Austin, Texas). The computer was controlled by MATLAB 7.5.0 and the Psychophysics Toolbox, Version 3.[18,19]

4.1.3 *Procedure*

This prospective study was reviewed and approved by the University of Texas, MD Anderson Cancer Center, institutional review board and was HIPAA compliant. The procedure for this study was the same as in Experiment 1, except that it had two types of testing, immediate and delayed. For the immediate testing, the test phases for each of two stimulus sets followed immediately after the study phase. The study and test phases were done back to back for one stimulus set before the subject went on to the next set. For the second and delayed type of testing, the test phases occurred 27 to 68 days (mean 49.9 days) after the study phases. In the study phases, each participant saw 72 images that were randomly taken from the 216 musculoskeletal radiographs or forest scenes. The study images were consecutively presented on the computer display, each for 3 s with no time between the images, resulting in a total time of 3 min and 36 s for the study phase. Participants were told to memorize the images in preparation for a recognition test. In the test phases, participants saw a sequence of 72 images, of which 36 were randomly chosen old images from the study phase, and the remaining 36 were completely new images.

For each participant, each image was randomly assigned to appear in the immediate or delayed test-timing group. Once assigned to immediate or delayed testing, images were then randomly selected to be in the study phase or the test phase or to be one of the images that appeared in both study and test phases. Therefore, in each of the two types of test timing (immediate and delayed) each observer saw a unique assortment of 108 images of each type of image (radiograph and scene), of which 36 would be seen only in the study phase, 36 only in the test phase, and 36 in both phases.

All participants completed the test and study phase blocks for the two image types and two test phase timings for a total of 288 test trials. The order of the blocks (radiographs versus forest scenes) was counterbalanced across participants. Our principal unit of analysis was the probability of a hit minus the probability of a false alarm (hits –false alarms), which we refer to as recognition accuracy. We also report performance in terms of percentage correct and assess differences using the signal detection measures of memory sensitivity (d′) and response criterion (c).

In this experiment, we also asked three radiologists to rank the musculoskeletal radiographs into three equal groups, those they thought would be easy to recognize, difficult to recognize, and of intermediate difficulty. All were allowed, indeed encouraged, to apply whatever criteria seemed appropriate to each individual to make this determination. These three radiologists had participated as observers. To decrease the likelihood that their experience with the images during their participation would affect their sorting, we waited 3 months between the end of data collection for these three and when they were given the images to sort.

4.2 *Results*

The aim of the second experiment was to further investigate the intentional visual memory of radiologists for the images of their expertise and natural scenes and test how the trace holds over time for the two different image types. We also wanted to add more heterogeneity into the radiographs by asking experts to memorize a more diverse group of musculoskeletal radiographs and at the same time introduce more homogeneity for real scenes by limiting the set only to exemplars from one category, forests. All of the statistical analysis was done on d' values.

4.2.1 *Immediate recall*

When comparing results on tests of immediate memory both in Experiments 1 and 2 with mixed model ANOVA, the findings show that increasing heterogeneity of a set of images from the domain of radiology expertise improved the radiologists' memory. The musculoskeletal radiographs (72%, s.e.m. = 1%; $d' = 1.48$) were remembered significantly better in comparison to the homogeneous set of chest radiographs [$F(1,21) = 97.9$, $p < .0001$; 65%, s.e.m. = 2.1%, $d' = 0.80$].

Our radiologists also recognized the musculoskeletal radiographs better than the forests [$F(1,10) = 116.74$, $p < .0001$; 67%, s.e.m. = 3.0%, $d' = 0.90$], yet not quite as well as radiologists in the first experiment had recognized the mixed-category natural scenes [$F(1,21) = 15.08$, $p < .001$; 85%, s.e.m. = 1.6%, $d' = 2.06$]. We saw the expected reverse pattern for natural scenes when we increased the homogeneity of the natural scene set to include only forest scenes. As those data imply, memory for forests alone (67%, s.e.m. = 3.0%, $d' = 0.90$) was significantly inferior [$F(1,21) = 97.9$, $p < .0001$] to recognition of a heterogeneous set of natural scenes composed of different scene categories (beach, mountain, cityscape, forest, room interior; 85%, s.e.m. = 1.6%, $d' = 2.06$) but did not differ from the homogeneous set of chest radiographs [$t(21) = –0.71$, $p = .49$; 65%, s.e.m. = 2.1%, $d' = 0.80$].

4.2.2 *Delayed recall*

A repeated measure ANOVA on data obtained in Experiment 2 which compares performance across two different testing times and image types shows that visual recognition memory for either type of image (musculoskeletal radiograph or forest) precipitously declines as the time between the study and test phases increases [$F(1,10) = 116.74$, $p < .00001$] [see Fig. (3)]. The decline of memory with time is worse for radiographs [$F(1,10) = 23.03$, $p < .001$], and when there is a delay between the study and test, the advantage in recognition for the more heterogeneous set of musculoskeletal radiographs in comparison to forests is erased (radiographs 50%, s.e.m. = 1.0%, $d' = 0.02$; forests 53%, s.e.m. = 2.0%, $d' = 0.16$), with both sets of images resulting in recognition accuracy similar to chance.

4.2.3 *Correlation with memorability scoring*

In Experiment 2, we also examined the possibility that radiologists might be able, based on their expertise, to predict which radiographs might be more recognizable than others. For the 216 musculoskeletal radiographs, the three radiologists that ranked the images agreed 76% of the time on their scores of the degree of difficulty for the individual images. Their rankings also positively correlate ($r = 0.55$ for all images; $r = 0.80$ for images with consensus) with the presence of an abnormality, with radiographs with no abnormalities being rated as more likely to be hard to recognize than those with abnormalities. For musculoskeletal images, we find that the rankings of radiologists significantly correlated with the readers' performance on the immediate recognition test (for all images $r = –0.23$ at $p < 0.05$, percentage

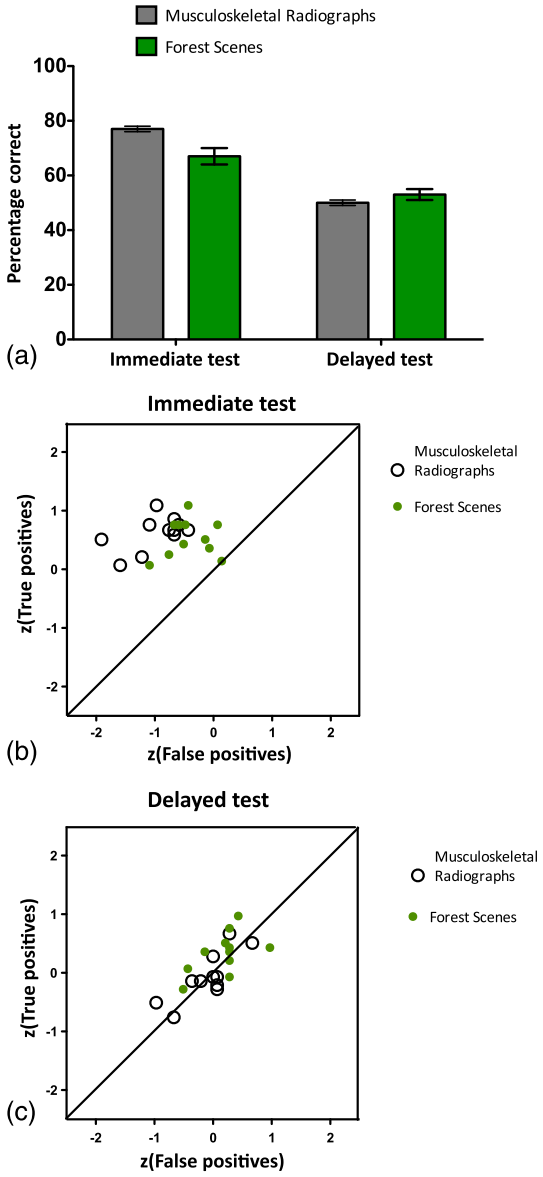

Fig. 3 Performance on visual recognition memory test of the radiologists on two images sets (forests and musculoskeletal radiographs) during the immediate and delayed recognition test. (a) Average accuracy for the two image sets across two different recognition times. Error bars represent standard error of mean. (b) Scatter plot of *z* score true positives against *z* score false positives by image set on immediate recognition test (average for forest scenes: 70% hits, 35% false positives; average for radiographs: 73% hits, 19% false positives). (c) Scatter plot of *z* score true positives against *z* score false positives by image set on delayed recognition test (average for forest scenes: 63% hits, 57% false positives; average for radiographs: 47% hits, 47% false positives).

of agreement on easy = 86%, medium = 77%, hard = 68%; for images of consensus $r = -0.36$ at $p < 0.01$, percentage of agreement on easy = 86%, medium = 76%, hard = 65%). Thus, the radiographs rated as easy to remember by radiologists were this time more easily remembered on the memory test and were likely to have abnormalities.

4.3 *Discussion*

Observer-performance experiments are commonplace in the diagnostic radiology literature. They can be designed in various ways depending on the question being asked. Generally, they have the common element that observers (usually radiologists) are asked to search out a particular feature on images in two or more different conditions. Sometimes the different conditions may be entirely separate types of images. In that situation, the observers' memory for the first set of images will not influence their interpretation of the second set because they are not seeing the same images again. For example, if one were testing the ability of radiologists to diagnosis meniscal tears on magnetic resonance imaging versus ultrasound, the appearance of the two modalities would be so different that one could reasonably expect that memory for the magnetic resonance imaging would not influence interpretation of the ultrasound and vice versa, if they are presented in an unconnected fashion.

At other times, the observers may be viewing the same images in different ways. For example, one experiment compared the observers' ability to find pulmonary nodules in different ambient noise conditions.[20] A big concern in such experiments is that memory for the first type of image viewed may affect the interpretation of the second type. Investigators use several different methods to try to decrease this effect, with one typical method being to allow a time gap between interpretations.

The concern that memory may affect interpretation on reexamination is warranted since research over the past decades on visual recognition memory has demonstrated that humans have a very large memory for visual information. Humans can correctly discriminate previously viewed photographs of scenes taken from everyday life versus new scenes with accuracy rates approaching or exceeding 90%, even when the foil photograph is from the same scene category[21] or when an interval up to 3 days has elapsed between viewings.[2] These experiments were done with heterogeneous image sets.

In observer performance studies, the image sets composed of radiographs are typically rather homogeneous. The images are usually presented in shades of gray and share a substantial number of elements, for example, two lungs, a heart, and 12 pairs of ribs on a chest radiograph. They differ only in smaller details.

So what happens when we directly compare memory of expert observers for a homogeneous set of representative images from their general domain of expertise to naïve observers and then contrast that performance to their memory for a variety of everyday scenes? Our findings replicate the results of Evans et al.[12] and show that radiologists' memory for chest radiographs, though significantly better than medically naïve observers', still does not come even close to their memory for everyday scenes. Even to radiologists, chest radiographs, though more scenes like than the mammograms used in an earlier study, with spatial layouts and less homogeneity of structures, are still not as memorable as everyday scenes.

Chest radiographs and real scene image sets differed markedly in their heterogeneity, so in order to test further how perceptual expertise as opposed to the heterogeneity or homogeneity of the image sets might modulate visual recognition memory, we tried to even the playing field between the two sets. Although different varieties of images could have been chosen, we believed that musculoskeletal radiographs and natural scenes restricted to forests would be reasonably comparable in terms of the amount of variety in the images. As it happened, radiologists' memory for musculoskeletal radiographs ($d' = 1.48$) was significantly better than for chest radiographs ($d' = .80$) and notably better than for images of forests ($d' = .90$), while their memory for the two

relatively homogeneous sets, chest radiographs and forests, was similar. Therefore, homogeneity or heterogeneity of the image sets seems to be of primary importance for memory, and expertise with the image set provides a lesser degree of advantage.

There is contradictory evidence regarding radiologists' memory for abnormal versus normal radiographs. In a study, which used both normal and abnormal chest radiographs carefully chosen to include only one example each of different types of abnormalities, senior staff radiologists remembered the abnormal images as well as they remembered human faces.[14] Another study of memory for radiographs was performed as part of an experiment testing radiologists' ability to distinguish two different positions of a central venous access catheter on frontal chest radiographs. In that study, the relevant abnormality, placement in the less desirable position, was present in half the images and did not improve recognition memory.[16] Both of our experiments were in the middle ground, both in terms of memory performance and in terms of variability of abnormalities. There were several different types of abnormalities, but there certainly were repetitions. We again believe this underscores the importance of variety in prompting memory.

With the chest radiographs, efforts to predict which images would be most easily remembered failed, while with the musculoskeletal set, memorability of radiographs was predictable, with their predictions correlating with the presence of an abnormality. One possible explanation for these results is that chest radiographs were too homogeneous and very hard to remember, thus the performance on those was close to floor, making the subtle effect of expertise and predictability hard to observe. Conversely, the musculoskeletal set was varied enough that performance was good, allowing us to observe how expertise with the images modulates memory for them. The details rendered meaningful to radiologists due to years of experience with radiographs introduced enough context in the image set to allow for radiographs to be more memorable than images of forests and their memorability predictable. Another possible reason why we find differing results when looking at the predictability of memorability of two different radiograph sets may relate to the subspecialization of the radiologists doing the ranking compared with those serving as observers. When chest radiologists ranked chest radiographs (and took as much time as they wished to consider each image), they may have noticed subtle findings that to them were interesting and would make the image memorable but that were not noticed by nonthoracic radiologists who were limited to 3 s to study each image. Conversely, when nonmusculoskeletal radiologists ranked musculoskeletal images, they brought to the task no greater level of training than the study participants, all but one of whom specialized in an area other than musculoskeletal imaging. Another possibility, of course, is that some lingering memory of their own performance as subjects may have guided some of the decisions of those ranking the musculoskeletal images, but we think that is unlikely given a 3-month gap. Though it is impossible to exclude possibility that an implicit impression about the images due to their own experience with the same could be contributing to the differences we observe between Experiments 1 and 2, it is possible that with a different method of ranking the memorability of the radiographs, correlation between memorability scoring and performance may have been achieved with the chest radiographs.

The second question we addressed was what happens to the memory for the images of expertise over time and how memory decays for them compare to everyday real scenes. Memory for both types of images came to chance levels after an average delay of 50 days between study and test phases. A similar decline has been reported in other studies where memory for photographs with miscellaneous real-world content already dropped to chance levels after 28 days post study.[22] Interestingly, we find that the decline is more severe for the images of expertise (i.e., radiographs) than for real-world images of forest. This decline is driven by a higher reduction in the hit rate rather than an increase in the rate of false alarms for radiographs in comparison to forests.

5 Conclusions

Memorability for both everyday scenes and images taken from a general domain of an observer's expertise (e.g., here radiology) is determined more by the variability of the set than the degree of experience of the observer with the type of image. Expertise with the images to be encoded into long-term pictorial memory allows for better encoding but not enough to make up for lack of idiosyncratic detail and homogeneity of the image set. It is not possible to reliably predict which images will be easily recognized. Nonetheless, for both homogeneous and heterogeneous image sets, idiosyncratic detail such as the presence of an abnormality can contribute to better recognition memory. Therefore, in reader-performance studies, avoiding the use of images with unique incidental abnormalities is recommended. Regarding the durability of pictorial memory, data indicate that visual recognition memory for both images of expertise and everyday scenes is not long. Prior studies have shown that with radiologic image sets even a gap of 1 to 3 days results in memory that is only slightly above chance, and we have here shown that memory for radiographs erodes to chance level within 7 weeks, suggesting that a time gap, though useful for avoiding interference from memory, does not need to be longer than one and a half months.

Acknowledgments

We would like to acknowledge the American Board of Radiology for making it possible to collect part of the data for this research. This research was supported in part by a grant from the John S. Dunn, Sr., Distinguished Chair in Diagnostic Imaging.

References

1. T. F. Brady, T. Konkle, and G. A. Alvarez, "A review of visual memory capacity: beyond individual items and toward structured representations," *J. Vision* **11**(5) (2011).
2. L. Standing, J. Conezio, and R. N. Haber, "Perception and memory for pictures: single-trial learning of 2500 visual stimuli," *Psychon. Sci.* **19**(2), 73–74 (1970).
3. M. A. Cohen, T. S. Horowitz, and J. M. Wolfe, "Auditory recognition memory is inferior to visual recognition memory," *Proc. Natl. Acad. Sci.* **106**(14), 6008–6010 (2009).
4. M. A. Cohen et al., "Auditory and visual memory in musicians and nonmusicians," *Psychon. Bull. Rev.* **18**(3), 586–591 (2011).
5. N. Charness, "Expertise in chess: the balance between knowledge and search," in *Toward a General Theory of Expertise: Prospects and Limits*, K. A. Ericsson and J. Smith, Eds., pp. 39–63, Cambridge University Press, Cambridge, England (1991).
6. W. G. Chase and H. A. Simon, "The mind's eye in chess," in *Visual Information Processing*, W. G. Chase Ed., pp. 215–281, Academic Press, New York (1973).
7. P. W. Frey and P. Adesman, "Recall memory for visually presented chess positions," *Memory Cognit.* **4**, 541–547 (1976).

8. F. Allard and J. L. Starkes, "Motor-skill experts in sports, dance and other domains," in *Toward a General Theory of Expertise: Prospects and Limits*, K. A. Ericsson and J. Smith, Eds., pp. 126–152, Cambridge University Press, Cambridge, England (1991).
9. J. M. Deakin and F. Allard, "Skilled memory in expert figure skaters," *Memory Cognit.* **19**, 79–86 (1991).
10. K. B. McKeithen et al., "Knowledge organization and skill differences in computer programmers," *Cognit. Psychol.* **13**, 307–325 (1981).
11. L. A. Hardesty et al., "Memory effect in observer performance studies of mammograms," *Acad. Radiol.* **12**, 286–290 (2005).
12. K. K. Evans et al., "Does visual expertise improve visual recognition memory?," *Atten. Percept. Psychophys.* **73**(1), 30–35 (2011).
13. A. Hillard et al., "The development of radiologic schemata through training and experience: a preliminary communication," *Invest. Radiol.* **20**, 422–425 (1985).
14. M. Myles-Worsley, W. A. Johnston, and M. A. Simons, "The influence of expertise on X-ray image processing," *J. Exp. Psychol.* **14**(3), 553 (1988).
15. G. R. Norman, L. R. Brooks, and S. W. Allen, "Recall by expert medical practitioners and novices as a record of processing attention," *J. Exp. Psychol.* **15**(6), 1166 (1989).
16. J. T. Ryan et al., "The 'memory effect' for repeated radiologic observations," *Am. J. Roentgenol.* **197**(6), W985–W991 (2011).
17. A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *Computational Visual Cognitive Laboratory at Mit Datasets*, http://cvcl.mit.edu (2001).
18. D. H. Brainard, "The psychophysics toolbox," *Spat. Vision* **10**, 433–436 (1997).
19. D. G. Pelli, "The VideoToolbox software for visual psychophysics: transforming numbers into movies," *Spat. Vision* **10**(4), 437–442 (1997).
20. P. C. Brennan et al., "The impact of acoustic noise found within clinical departments on radiology performance," *Acad. Radiol.* **15**(4), 472–476 (2008).
21. T. Konkle et al., "Scene memory is more detailed than you think the role of categories in visual long-term memory," *Psychol. Sci.* **21**(11), 1551–1556 (2010).
22. R. S. Nickerson, "Short-term memory for complex meaningful visual configurations: a demonstration of capacity," *Can. J. Psychol.* **19**(2), 155 (1965).

Karla K. Evans is assistant professor of psychology and neuroscience at University of York. Her current research interest revolves around visual awareness, visual episodic memory, perceptual expertise, and medical image perception.

Edith M. Marom is a professor of radiology at the University of Tel Aviv, Sheba Medical Center, Israel, as well as at the University of Texas M. D. Anderson Cancer Center in Houston, Texas. Her research interests include accuracy of oncological chest imaging with CT, FDG PET-CT, and MRI, interpretation methods used by radiologists, and memory for medical images.

Myrna C. B. Godoy is assistant professor of radiology at the University of Texas M. D. Anderson Cancer Center in Houston, Texas. Her research interest are lung cancer screening and biomarker validation, lung cancer staging and tumor response assessment, pulmonary nodule characterization and management, pulmonary infection in the oncologic setting, dual-energy computed tomography imaging; iterative image reconstruction and radiomics.

Diana Palacio is a cardiothoracic, pediatric radiologist, and associate professor in the Department of Medical Imaging of the University of Arizona. Her research interests are imaging of chest and heart diseases in adults and general pediatric imaging.

Tara Sagebiel is an assistant professor of radiology at the University of Texas M. D. Anderson Cancer Center in Houston, Texas. Her academic interests include medical quality improvement, appendiceal cancer imaging, gastric cancer imaging, and memory for medical images.

Sonia Betancourt Cuellar is a assistant professor, Division of Diagnostic Imaging, Department of Diagnostic Radiology, The University of Texas MD Anderson Cancer Center, Houston, TX, and adjunct assistant professor appointment in Division of Diagnostic Imaging, The University of Texas Health Science Center at Houston, Houston, TX. Completed radiology residency in Bogota, Colombia, and thoracic radiology fellowship training at MD Anderson Cancer Center. Her research interests are imaging evaluation in esophageal carcinoma, biomarkers in esophageal carcinoma, and correlation with PET/CT.

Mark McEntee is associate professor of medical radiation science. His current research interests revolve around perception in medical imaging and performance errors, as well as radiation dose and image quality analysis with the aim of improving radiological diagnostic performance.

Charles Tian recently completed his undergraduate studies at the University of Chicago. He was a summer student at M. D. Anderson Cancer Center for two summers, one of which was spent working with Dr. Tamara Miner Haygood. During that time, he assisted with her project on radiologists' memory of features of chest radiographs, a good portion of data from which is included in this manuscript.

Patrick C. Brennan is leader of the Imaging, Optimisation and Perception Group at the University of Sydney. He has presented at the major international imaging meetings and published over 170 original papers. His work is having an important impact on clinical practice across the globe. He has won 2 medals of excellence for teaching and has acted as undergraduate, graduate, and PhD examiner in 9 universities across Europe, Asia, and the Americas.

Tamara Miner Haygood is associate professor of radiology at the University of Texas M. D. Anderson Cancer Center in Houston, Texas. Her research interests include interpretation methods used by radiologists, interpretation efficiency, and memory for medical images.